



VERIFICATION OF TRANSLATION

I, the below named translator, hereby declare:

that my name and my post office address are as stated below; and

that I am knowledgeable in the English and Korean languages and that I believe the following is a true and complete translation into the English language of Korean Patent Application No. 1995-53941 filed in the Korean Patent Office on the 22nd day of December, 1995 for Letters Patent, including a true translation of the Official Certificate of the Application.

Signed this 21st day of September 1998

JEONG SOOK YI

Full name of translator

Jeong S. Yi

Signature of translator

HAECHEON BLDG. 2F 741-40, YEOKSAM-1 DONG,
KANGNAMKU, SEOUL, 135-081

Post Office Address

Received

OCT 21 1998

Group 2700



SPECIFICATION

1. TITLE OF THE INVENTION

Phoneme Dividing Method Using Multilevel Neural Network

2. BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a system to which the present invention is applied.

FIG. 2 shows a configuration of a multilevel neural network used for the present invention.

FIG. 3 is a flickered of one embodiment of the present invention.

* description of the principle reference numerals

1: voice input portion

2: preprocessor

3: MLP phoneme dividing portion

4: phoneme border output portion

3. DETAILED DESCRIPTION OF THE INVENTION

The present invention relates to a phoneme dividing method using a multilevel neural network.

Conventional phoneme dividing technologies complicate their systems by finding the border of phonemes through an analysis to which prefixed various phonetic knowledge and rules after extracting the frequency component, that is, the spectrogram, from an acoustic signal.

Without an effective and optimal method for combining various knowledge and rules used in phoneme division, the performance of system is not reliable and drastically deteriorated depending upon the change of situation.

There is a method for finding the border of phoneme by comparing characteristic patterns with an incoming signal in phoneme division after previously extracting the characteristics of all phonemes and storing them in patterns. This method requires information on the characteristic patterns for all phonemes to undesirable increase the volume of memory of the system and also the amount of calculation in performance.

Therefore, it is an object of the present invention to provide a phoneme dividing method using a multilevel neural network for precisely and efficiently capturing the point of phoneme border, using only the variation of vocal signal appearing at the border of phonemes, without additional knowledge for phoneme itself, to be thereby utilized in application fields requiring automatic phoneme division or phoneme labeling.

To accomplish the object of the present invention, there is

provided a phoneme dividing method using a multilevel neural network applied to a phoneme dividing apparatus having a voice input portion for outputting a vocal sample digitally converted from voice made, a preprocessor for extracting a characteristic vector suitable for phoneme division, from the vocal sample input from the voice input portion, a multi-layer perception(MAP) phoneme dividing portion for finding and outputting the border of phoneme, using the characteristic vector of the preprocessor, and a phoneme border outputting portion for outputting position information on the border of phoneme of the MAP phoneme dividing portion in the form of frame position, the method comprising the steps of: (a) sequentially segmenting and framing voice with digitalized voice samples, extracting characteristic vectors by vocal frames, and extracting an inter-frame characteristic vector of the difference between nearby frames of the characteristic vectors by frames, to thereby normalize the maximum and minimum of the characteristics; (b) initializing weights present between an input layer and hidden layer and between the hidden layer and output layer of the MAP, designating an output target data of the MAP, unpitying the characteristic vectors to the MAP for learning, and storing and finishing information on the weight obtained through learning and the standard of the MAP if the reduction rate of mean squared error converges within a permissible limit; and (c) reading the weight obtained in the step (b), receiving the characteristic vectors,

performing an operation of phoneme border discrimination to generate an output value, discriminating the phoneme border according to the output value, and if the current analyzed frame arrives two frames preceding the final frame of incoming voice, outputting a frame number indicative of the border of phoneme as a final result.

Hereinafter, a preferred embodiment of the present invention will be described below.

In FIG. 1, reference numeral 1 represents a voice input portion. Reference numeral 2 is a preprocessor, 3 being a multi-layer perception (MAP) phoneme dividing portion, and 4 being a voice border output portion.

Voice input portion 1 comprises a microphone for converting an aerial vocal waveform into an electric vocal signal, a band-pass filter for eliminating low-frequency noise and high-frequency aliasing from the vocal signal input as an electric analog signal, and an analog-to-digital converter (ADC) for converting the analog vocal signal into a digital vocal signal. The voice input portion output a vocal sample converted into digital from the voice, to preprocessor 2.

Preprocessor 2 extracts characteristic vectors suitable for phoneme division from the vocal samples input from voice input portion 1, and outputs them to MLP phoneme dividing portion 3. MLP phoneme dividing portion 3 finds the border of phoneme, using

characteristic vectors input from preprocessor 2, and outputs the result to phoneme border output portion 4. Phoneme border output portion 4 outputs position information on phoneme border automatically divided in MLP phoneme dividing portion 3 in the form of frame position.

Referring to FIG. 2, one embodiment of the present invention implements an effective and reliable automatic phoneme segmenter by using a multi-layer perceptron (MLP), one kind of neural network, in order to complement the drawbacks of the conventional phoneme dividing method based upon knowledge or rules.

A phoneme dividing method using MLP is very favorable to solving decrease of performance caused due to imperfect modeling of knowledge or rules on the border of phoneme contained in a vocal signal. In this method, functions required in phoneme division are learned voluntarily from the characteristic vectors extracted from a large amount of vocal data so that the MLP itself finds the knowledge or rules contained in the vocal signal, without previously introducing specific suppositions, rules or knowledge on the border of phoneme. Accordingly, the method of the present invention eliminates the introduction of unsure supposition or additional processing of distribution or modeling of the vocal signal in order to facilitate its modeling.

MLP used in the present invention is made in a multiple structure of three layers of input, hidden and output layers. As

shown in the drawing, the input layer placed on the bottom is made with 73 input nodes of 72 input nodes for inter-frame characteristic vectors extracted from four inter-frame differences generated among five sequential frames, and one input node for an input value 1 to be used instead of the threshold value comparison process in the hidden layer of MLP.

The output node of the output layer is made with two nodes of the first node indicative of the border of phoneme, and the second node not indicative of the border of phoneme. The hidden layer placed between the input and output layers is to perform nonlinear discrimination that the MLP must implement actually.

The following nonlinear sigmoid function is used for the activation function of the hidden layer.

$$y = (\exp(x) - 1) - (\exp(x) + 1)$$

where x and y represent the input and output of the activation function, respectively.

The number N of nodes of the hidden layer is known to be closely relevant to the final function of MLP. It is noted through an experiment using various kinds of data that it is appropriate that the number of nodes be between 10 and 30. Between the input layer and hidden layer and between the hidden layer and output layer, there are weights which connect all the nodes of the respective layers. Because the weights connect all the nodes between the layers, its number is $73 \times N$ (the number of input nodes

x the number of hidden nodes) in case of the input layer and hidden layer. The number of weights is $N \times 2$ (the number of hidden node x the number of output node). These functions are previously obtained through learning using an error back propagation algorithm, stored
5 in a memory, and then read out in phoneme division.

FIG. 3 shows a procedure of the phoneme division algorithm in preprocessor 2 and MLP phoneme dividing portion 3, having two parts of learning process and dividing process of the MLP phoneme dividing algorithm.

10 Above all, the process of voice framing and characteristic vector extraction is performed in preprocessor 2 and used commonly in the learning and dividing processes. In selection the characteristic vectors in the present invention, factors explicitly indicative of the difference of spectrum between frames are induced
15 in order to use the fact that the variation of vocal spectrum is severe at the border phonemes.

Voice is sequentially segmented in a length so long as to extract the characteristics of voice from digitalized voice samples, for the purpose of voice framing in step 10. Voice framing
20 is performed by taking Hamming windows in the length of 16 msec every 10 msec with respect to the overall incoming vocal samples.

Then, the characteristic vectors are extracted from the voice frames in step 11 containing two substeps. In the first step, characteristic vectors by frames effectively indicative of the

characteristics of voice are extracted on basis of phonetic knowledge, with respect to the respective voice frames obtained before. In the second step, inter-frame characteristic vectors of the difference between nearby frames with respect to the characteristic vectors by frames obtained in the first step are extracted to be used as the final characteristic vectors input to MLP phoneme dividing portion 3.

For more detailed description of the above procedure, the characteristic vectors first obtained with respect to the respective frames are as follows.

(1) frame energy: indicates the intensity of phonation by frames and is found according to the following equation.

$$ENG_FRM(t) = \log_{10} \left(\sum_n s(n) * s(n) \right), n=0,1,\dots,N$$

where $s(n)$ represents a vocal sample belonging to the t_{th} frame, and N represents the length of vocal frame.

(2) 16th degree Mel-scaled fast Fourier transform (FFT):

First, FFT is performed in order to obtain the spectrum, the frequency characteristic of voice by frames, and the frequency component of voice is classified into predetermined 16 frequency bands similar to the human hearing characteristics, to obtain the degree energy by bands which is used as the coefficient of the Mel-

scaled FFT. The j_{th} degree Mel-scaled FFT coefficient for frame index t is obtained as follows.

$$MSFC(j, t) = \log_{10} \left(\sum_{f=1}^{16} s(j, t, f) \right)$$

5

where f represents the frequency belonging to the respective frequency bands;

j is the index of the respective frequency bands; and $s(j, t, f)$ is j_{th} degree frequency band amplitude spectrum of t_{th} frame obtained from FFT by frequencies.

10

(3) energy ratio by bands: It is very important to precisely discriminate phoneme into voiced sound and voiceless sound in phoneme division. The difference between voiced and voiceless sounds is the distribution of energy by frequency bands. In order to discriminate voiceless and voiced sounds in the present invention, the low-frequency energy between 0 and 3 kHz and the high-frequency energy between 3 and 8 kHz are obtained respectively, and their ratio is selected as one of the characteristic vectors.

15

20

$$\begin{aligned}
ENG_RTO(t) &= \log_{10}(ENG_LOW(t)) - \log_{10}(ENG_HIGH(t)) \\
ENG_LOW(t) &= \sum_f s(f, t), f=0, \dots, 3kHz \\
ENG_HIGH(t) &= \sum_f s(f, t), f=3, \dots, 8kHz
\end{aligned}$$

Where $ENG_LOW(t)$, and $ENG_HIGH(t)$ are energies of the low and high frequency bands of the t_{th} voice frame, respectively, which are
5 obtained by the sum of components contained in the respective bands at the amplitude spectrum obtained in the FFT.

The inter-frame characteristic vectors used as the final input of MLP phoneme dividing portion 3 can be obtained by finding the difference between nearby frames with respect to the first
10 characteristic vectors by frames on basis of the fact that the variation of phoneme division occurs at the border of phonemes.

(1) difference of frame energy between nearby frames

$$dENG_FRM(t) = ENG_FRM(t) - ENG_FRM(t-1)$$

(2) inter-frame difference of 16_{th} degree Mel-scaled FFT

$$dMSFC(j, t) = MSFC(j, t) - MSFC(j, t-1) \quad , \quad J=0, 1, \dots, 15$$

Here, j represents the respective degrees of the coefficients.

(3) inter-frame difference of energy ration by frames

$$dENG_RTO(t) = ENG_RTO(t) - ENG_RTO(t-1)$$

After the characteristic vectors are extracted as above, they

are normalized in step 12 whose maximum and minimum become 1 and -1, respectively, in order to be used as the input of MLP phoneme dividing portion 3.

In the learning process of MLP phoneme dividing portion 3 using the normalized characteristic vectors, weights present between the input and hidden layers and the hidden and output layers are initialized in step 13 as the initial learning step of MLP phoneme dividing portion 3. The initial value is established as an arbitrary value distributed between 1 and -1.

After this step, output target data of the output layer, which teaches finding the border of phonemes, is designated in step 14. The output target data by frames is equal to the number of the MLP output nodes, having values of (1,-1) in case of the border of phoneme and (-1,1) in other cases. This output target data is made to coincide with the frame position of corresponding characteristic vectors using information on the border of phoneme obtained from previously phoneme-divided voice database.

After the designation of output target data, the characteristic vectors, learning data, are input to the input layer of the MLP in step 15 so as to teach the MLP in step 16. The input layer has 73 nodes of 72 input nodes for the input of the four sequential inter-frame characteristic vectors and one input node for 1 to be input instead of the threshold value comparison procedure of the hidden layer.

The four inter-frame characteristic vectors are extracted among four intervals generated from five frames including preceding and succeeding two frames $t-2$, $t-1$, $t+1$, $t+2$, centering on the currently analyzed frame t , as shown in the lower portion of FIG.

5 2. The learning algorithm of the phoneme dividing MLP uses the generally used error back propagation algorithm.

After this learning process of MLP, if the reduction rate of mean squared error converges within a permissible limit in step 17, the weights obtained through learning and information on the standard of the MLP are stored in step 18 to finish the learning process. After the learning process, the voice is sequentially segmented in a length so long as to extract the characteristics of voice from the digitalized vocal samples for voice framing in step 10, and the characteristic vectors are extracted in step 11 and
15 normalized in step 12.

The weights obtained in the learning process are read into the hidden layer of the MLP in step 19. Then, the 72 characteristic vectors obtained in the above process are input in the sequence of the input nodes of the MLP, and 1 is input to the final 73_{th} input
20 node in step 20.

In MLP phoneme dividing portion 3, the output value for phoneme border discrimination is produced through the following MLP operation with respect to incoming characteristic vectors in step 21.

$$ID(j) = SGMOD \sum_i IN(i) \times WGT_IH(i, j), i=0, 1, \dots, 72 \quad j=0, 1, \dots, N-ID(N-1)=1,$$

$$UT(k) = SGMOD \sum_j HID(j) \times WGT_HO(j, k), j=0, 1, \dots, N-1, k=0, 1$$

Where $IN(j)$ represents the input of the i_{th} input node;

5 $OUT(k)$ is the output of the k_{th} output node;

$WGT_IH(i, j)$ is the weight connecting the i_{th} input node and j_{th} hidden node;

$WGT_HO(j, k)$ is the weight connecting the j_{th} hidden node and the k_{th} output node; and

10 $SGMOD$ represents the aforementioned sigmoid function.

Value 1 is designated to the final hidden node instead of the threshold comparison procedure in the final output node.

When the output values operated in MLP phoneme dividing portion 3 are compared in discriminating the border of phoneme, if
15 the first output value $OUT(0)$ is positive, the analyzed frame is the border of phoneme. In contrast, if $OUT(1)$ is positive, it is determined in step 22 that the frame is not the border of phoneme.

In step 23, it is checked whether the currently analyzed frame arrives two frames preceding the final frame of the incoming voice.
20 If not, the procedure of inputting the characteristic vectors to the MLP input layer is iterated. If the currently analyzed frame arrives two frames preceding the final frame, the value expressed as a frame number indicative of the border of phoneme is output as the final result in step 24, and the whole procedure ends.

In implementing a voice recognition system which makes it possible the conversation between human beings and machines, the present invention operating as above divides voice in units of phoneme and enables precise and effective phoneme division preprocessing essentially required phoneme recognition based upon phoneme division with respect to the divided phoneme segments. In addition, the present invention enables automatic voice experts in constructing a large volume of phoneme-divided voice database required in implementing a phoneme-unit voice recognition and voice mixing system. This reduces time and cost.

Although the present invention has been described above with reference to the preferred embodiments thereof, those skilled in the art will readily appreciate that various modifications and substitutions can be made thereto without departing from the spirit and scope of the invention as set forth in the appended claims.

4. CLAIMS

1. A phoneme dividing method using a multilevel neural network applied to a phoneme dividing apparatus having a voice input portion for outputting a vocal sample digitally converted from voice made, a preprocessor for extracting a characteristic vector suitable for phoneme division, from the vocal sample input from the voice input portion, a multi-layer perceptron (MLP) phoneme dividing portion for finding and outputting the border of phoneme, using the characteristic vector of the preprocessor, and phoneme border outputting portion for outputting position information on the border of phoneme of the MLP phoneme dividing portion in the form of frame position, said method comprising the steps of:

(a) sequentially segmenting and framing voice with digitalized voice samples, extracting characteristic vectors by vocal frames, and extracting an inter-frame characteristic vector of the difference between nearby frames of the characteristic vectors by frames, to thereby normalize the maximum and minimum of said characteristics;

(b) initializing weights present between an input layer and hidden layer and between the hidden layer and output layer of said MLP, designating an output target data of said MLP, inputting said characteristic vectors to said MLP for learning, and storing and finishing information of the weight obtained through learning and

the standard of said MLP if the reduction rate of mean squared error converges within a permissible limit; and

(c) reading the weight obtained in said step (b), receiving said characteristic vectors, performing an operation of phoneme border discrimination to generate an output value, discriminating the phoneme border according to the output value, and if the current analyzed frame arrives two frames preceding the final frame of incoming voice, outputting a frame number indicative of the border of phoneme as a final result.

10

2. The method as claimed in claim 1, wherein the voice framing of said step (a) is performed by taking a Hamming Window in a length of 16 msec every 10 msec with respect to the overall incoming vocal samples.

15

3. The method as claimed in claim 1, wherein the phoneme border discrimination of said step (c) is performed in such a manner that output values generated through operation are compared, and then it is determined that if output value OUT(0) is positive, an analyzed frame is the border of phonemes, and if output value OUT(1) is positive, the frame is not the border of phonemes.

20

ABSTRACT

A phoneme dividing method using a multilevel neural network applied to a phoneme dividing apparatus having a voice input portion, a preprocessor, a multi-layer perception (MLP) phoneme dividing portion, and a phoneme border outputting portion includes the steps of: (a) sequentially segmenting and framing voice with digitalized voice samples, extracting characteristic vectors by vocal frames, and extracting an inter-frame characteristic vector of the difference between nearby frames of the characteristic vectors by frames, to thereby normalize the maximum and minimum of the characteristics; (b) storing information on the weight obtained through learning and the standard of the MLP; and (c) reading the weight obtained in the step (b), receiving the characteristic vectors, performing an operation of phoneme border discrimination to generate an output value, discriminating the phoneme border according to the output value, and if the current analyzed frame arrives two frames preceding the final frame of incoming voice, outputting a frame number indicative of the border of phoneme as a final result.

FIG. 1

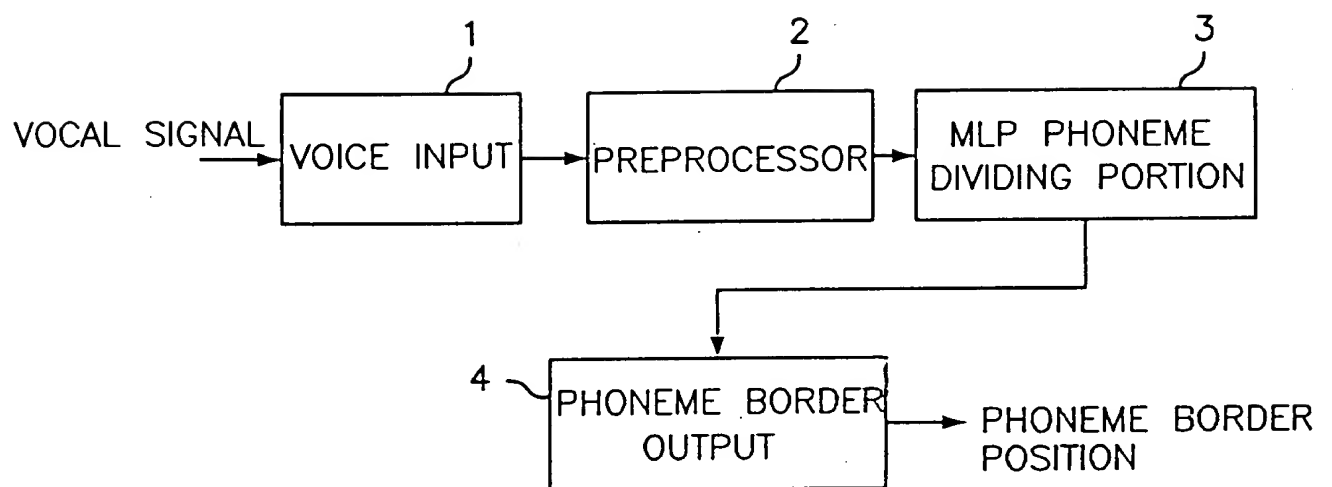


FIG. 2

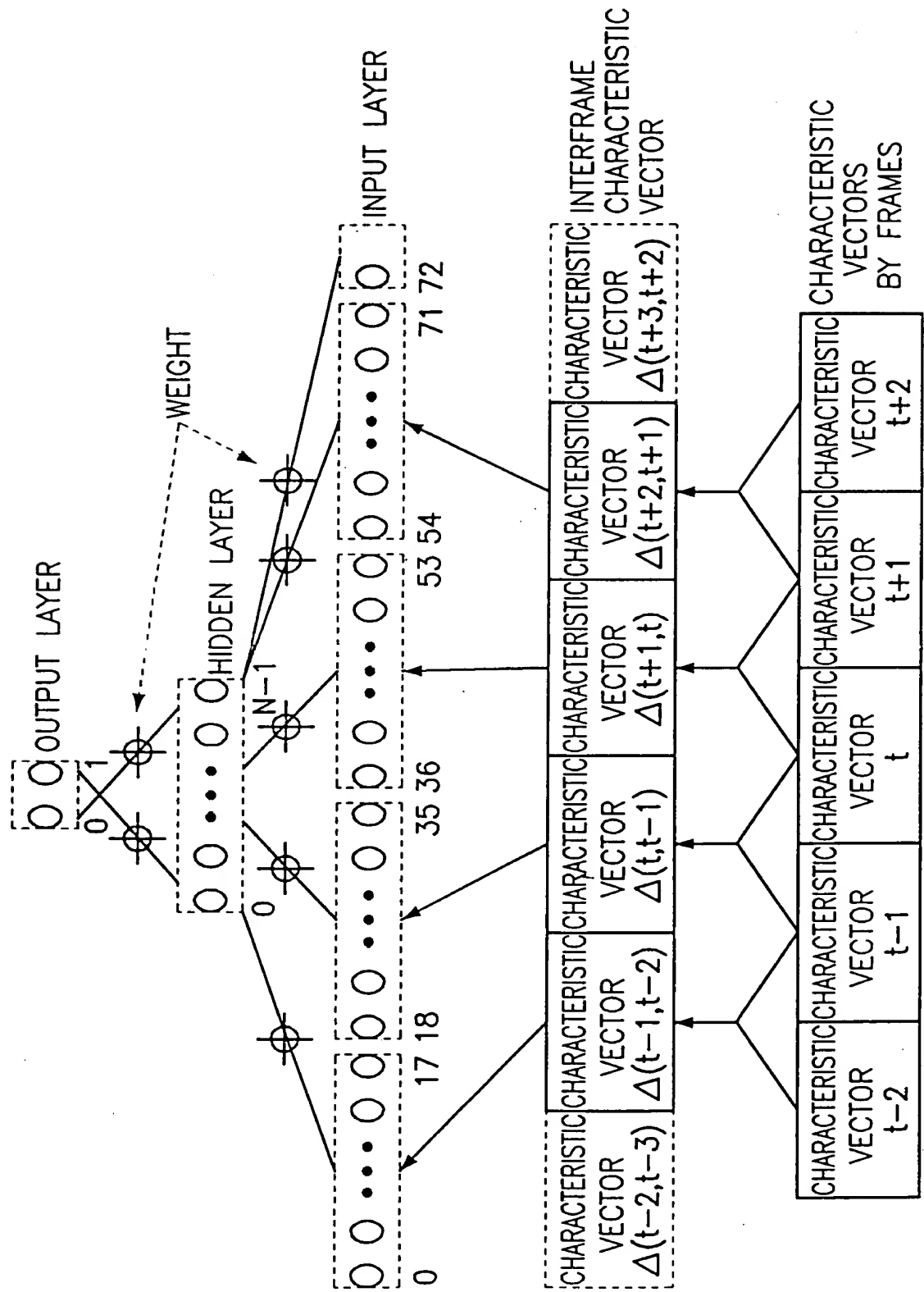
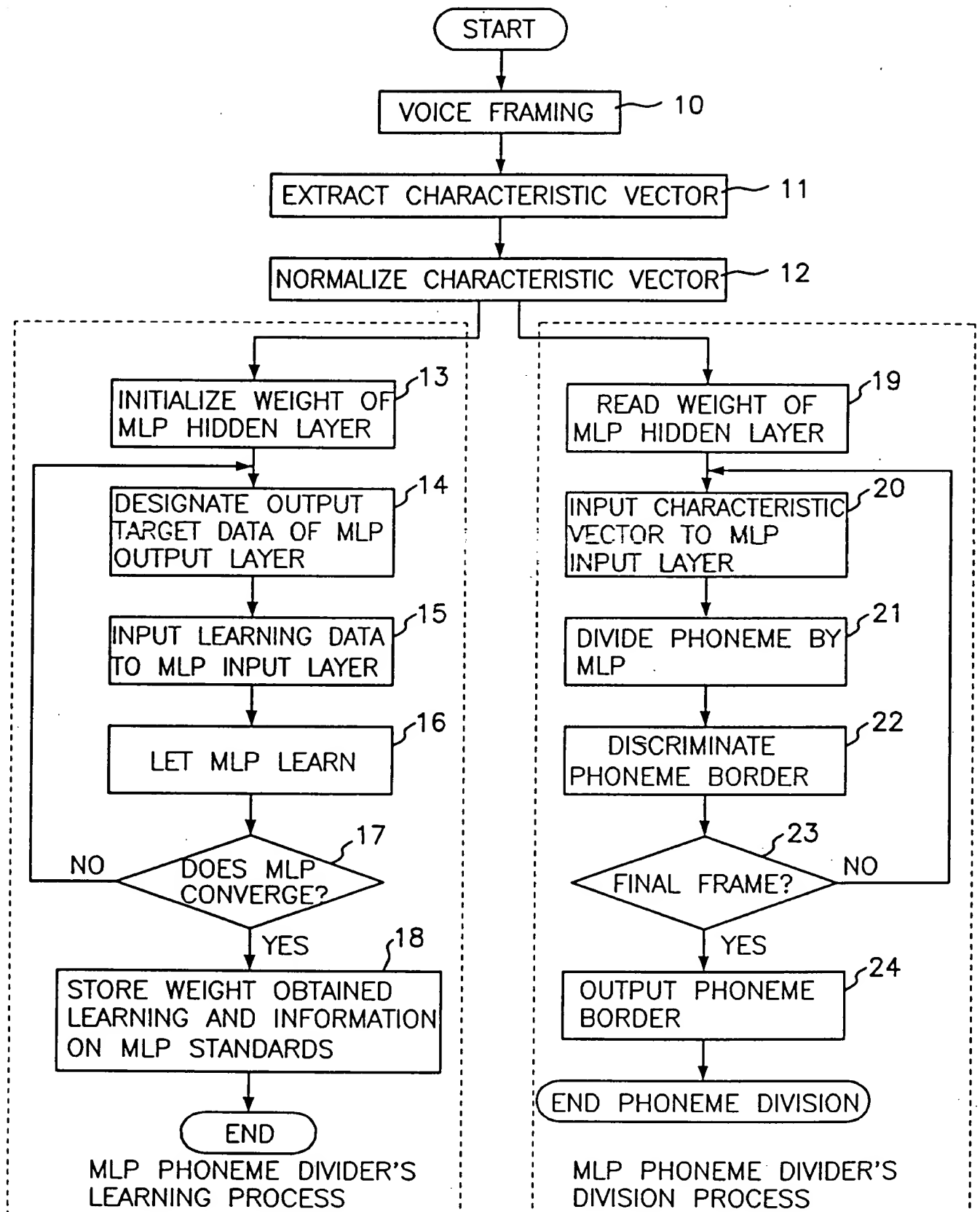


FIG. 3



DECLARATION AND POWER OF ATTORNEY FOR PATENT APPLICATION

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name, I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled:

PHONEME DIVIDING METHOD USING MULTILEVEL NEURAL NETWORK

the specification of which (check one)

☐

is attached hereto.

☒

was filed on November 19, 1996

as Application Serial No. 08/746,981

and was amended on _____

(if applicable)

I hereby state that I have reviewed and understand the contents of the above-identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose information which is material to the examination of this application in accordance with Title 37, Code of Federal Regulations, §1.56(a).

I hereby claim foreign priority benefits under Title 35, United States Code, §119 of any foreign application(s) for patent or inventor's certificate listed below and have also identified below any foreign application for patent or inventor's certificate having a filing date before that of the application on which priority is claimed:

Prior Foreign Application(s)

Priority Claimed

<u>1995-53941</u> (Number)	<u>Republic of Korea</u> (Country)	<u>22/12/1995</u> (Day/Month/Year Filed)	<input checked="" type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	<input type="checkbox"/> Yes	<input type="checkbox"/> No

I hereby claim the benefit under Title 35, United States Code, §120 of any United States application(s) listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States application in the manner provided by the first paragraph of Title 35, United States Code, §112, I acknowledge the duty to disclose material information as defined in Title 37, Code of Federal Regulations, §1.56(a) which occurred between the filing date of the prior application and the national or PCT international filing date of this application:

_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status: patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status: patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status: patented, pending, abandoned)
_____ (Application Serial No.)	_____ (Filing Date)	_____ (Status: patented, pending, abandoned)

(Continued on Page 2)

I hereby appoint as principal attorneys: Donald R. Antonelli, Reg. No. 20,296; David T. Terry, Reg. No. 20,178; Melvin Kraus, Reg. No. 22,466; William I. Solomon, Reg. No. 28,565; Gregory E. Montone, Reg. No. 28,141; Ronald J. Shore, Reg. No. 28,577; Donald E. Stout, Reg. No. 26,422; Alan E. Schiavelli, Reg. No. 32,087; James N. Dresser, Reg. No. 22,973 and Carl I. Brundidge, Reg. No. 29,621 to prosecute and transact all business connected with this application and any related United States application and international applications. Please direct all communications to the following address:

ANTONELLI, TERRY, STOUT & KRAUS, LLP
Suite 1800
1300 North Seventeenth Street
Arlington, Virginia 22209
Telephone: (703) 312-6600
Fax: (703) 312-6666

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further, that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

(Full Name)

(Signature)

Date	September 7, 1998	Inventor	YOUNG JIK LEE	<i>Young Jik Lee</i>
Residence	Daejeon, Korea	Citizenship	Republic of Korea	
Post Office Address	Hanbit Apartment #111-601, Uheun-dong, Yuseong-ku, Daejeon, Korea			
Date	September 7, 1998	Inventor	YOUNG JOO SUH	<i>Young Joo Suh</i>
Residence	Kyongsangbuk-do, Korea	Citizenship	Republic of Korea	
Post Office Address	#357 Eonha-2dong, Yeongcheon-shi, Kyongsangbuk-do, Korea			
Date	September 7, 1998	Inventor	JAE WOO YANG	<i>J. W. Yang</i>
Residence	Daejeon, Korea	Citizenship	Republic of Korea	
Post Office Address	Hanbit Apartment #106-1005, Uheun-dong, Yuseong-ku, Daejeon, Korea			
Date		Inventor		
Residence		Citizenship		
Post Office Address				
Date		Inventor		
Residence		Citizenship		
Post Office Address				
Date		Inventor		
Residence		Citizenship		
Post Office Address				
Date		Inventor		
Residence		Citizenship		
Post Office Address				
Date		Inventor		
Residence		Citizenship		
Post Office Address				
Date		Inventor		
Residence		Citizenship		
Post Office Address				